36-709: Advanced Statistical Theory I **Spring 2019**

Lecture 18: March 28

Lecturer: Alessandro Rinaldo *Scribe: Chirag Gupta*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In Section 18.1, we will consider selection consistency (also known as *sparsistency*) for the Lasso. In Section 18.2, we will introduce oracle inequalities.

18.1 Sparsistency for Lasso

In this section we continue analyzing the Lasso estimator for linear regression. Because the Lasso behaves like a soft-thresholding operator, it returns sparse solutions. This motivates us to ask the question if Lasso recovers the right support.

Goal: Estimate $S := \text{supp}(\beta^*)$ exactly. This is generally very hard.

Define $\mathbf{X}_S$ as the sub-matrix of $\mathbf{X}$ composed only by the column belonging to an index set S. Then suppose the following assumptions hold:

1. **Smallest eigenvalue:**
$$\lambda_{\min}\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right) \geq c_{\min} > 0.$$

2. **Incoherence:**
$$C_{\mathbf{X}} = \left|\left|\left|\mathbf{X}_{S^c}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\right|\right|\right|_\infty := \max_{j\in S^C}\left\|\mathbf{X}_j^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\right\|_1 \geq c_{\min} > 0.$$

Theorem 18.1 *Set $\lambda_n \geq \frac{2}{1-\alpha}\left\|\mathbf{X}_{S^c}^T\Pi_{S^\perp}\left(\frac{\epsilon}{n}\right)\right\|_\infty$, where $\Pi_{S^\perp} := \mathbf{I} - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T$. If the assumptions above are true, then*

1. *The Lasso solution is unique.*
2. *$\hat{S} := \{j : \hat{\beta}_j^{LASSO} \neq 0\} \subset S$.*
3. *$\left\|\hat{\beta}^{LASSO} - \beta^*\right\|_\infty \leq B_n(\lambda_n, \mathbf{X}) := \left\|\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right)^{-1}\left(\frac{\mathbf{X}_S^T\epsilon}{n}\right)\right\|_\infty + \lambda_n\left|\left|\left|\left(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}\right)^{-1}\right|\right|\right|_\infty.$*

If in addition, $\min_{j \in S} \left|\beta_j^*\right| > B_n(\lambda_n, \mathbf{X})$*, then* $\hat{S} = S$*.*

Remark: If for some $v > 0$, $C_{\mathbf{X}} \geq 1 + v$, then with probability at least 0.5, there is no model selection consistency.

Proof technique. See Martin Wainwright's seminal paper for the proof [MW06]. We skip the proof in this class. The proof technique used is called the **Primal-dual witness construction**.

18.2 Oracle inequalities

Oracle inequalities are a framework to compare bounds on error rates of an estimator to an oracle estimator. An oracle estimator is typically either an unknown optimal estimator, or it makes use of additional information not known to the algorithm. In this section we introduce oracle inequalities for regression.

18.2.1 Oracle inequalities for regression

We observe n pairs $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$, where Y_i's are independent random variables and $\mathbf{x}_i$'s are fixed (non-random) points in $\mathbb{R}^d$ such that for every i,

$$Y_i = f^*(\mathbf{x}_i) + \epsilon_i,$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Note that for linear regression we assumed $f^*(\mathbf{x}_i) = \mathbf{x}_i^T \beta$ for some β.

Suppose we have a dictionary of functions on $\mathbb{R}^d$: $D = \{f_1, \ldots f_m\}$. The goal is to estimate f^* using some linear combination of functions in D: $\hat{f}(\cdot) = \sum_{j=1}^m \hat{\theta}_j f_j(\cdot)$ for some $(\theta_1, \ldots, \theta_m) \in K \subset \mathbb{R}^m$.

Remark. If $m = d$ and $f_j(\mathbf{x}) = \mathbf{x}_j$, then $f(\mathbf{x}) = \sum_{j=1}^m \theta_j f_j(\mathbf{x}) = \theta^T \mathbf{x}$. Observe that this is the same as linear regression.

Now, for any estimator $\hat{f}$, define

$$R(\hat{f}) := \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \left(\hat{f}(\mathbf{x}_i) - f^*(\mathbf{x}_i)\right)^2\right] = \mathbb{E}\left[\text{MSE}(\hat{f})\right] = \frac{\mathbb{E}\left[\left\|\underset{\sim}{\hat{f}} - \underset{\sim}{f^*}\right\|_2^2\right]}{n}.$$

where in the last term we define $\underset{\sim}{\hat{f}}$ as the column vector in $\mathbb{R}^n$ with the i'th entry equal to $\hat{f}(\mathbf{x}_i)$. Note that if f is a fixed non-random linear combination of functions in D, then

$$R(f) = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2.$$

We first consider an oracle defined as follows:

- For $\theta \in K \subset \mathbb{R}^m$ define,
$$f_\theta(\cdot) = \sum_{j=1}^{m} \theta_j f_j(\cdot).$$
- Then define f_{ORACLE} such that it satisfies,
$$R(f_{\text{ORACLE}}) = \inf_{\theta \in K} R(f_\theta).$$
Note that none of the entities in the above definition are stochastic. Thus we can say:
$$\text{MSE}(f_{\text{ORACLE}}) = \inf_{\theta \in K} \text{MSE}(f_\theta).$$
- **Remark.** $f^* \neq f_{\text{ORACLE}}$, unless f^* can be represented as linear combinations in D.

Our target is to produce an estimator $\hat{f}$ such that $R(\hat{f})$ is as close as possible to $R(f_{\text{ORACLE}})$. An estimator $\hat{f}$ satisfies an oracle inequality if,

$$R(\hat{f}) \leq C \cdot R(f_{\text{ORACLE}}) + T(n, D, f^*, K, d),$$

where $C \geq 1$ and T is vanishing in N. The inequality is said to be sharp if $C = 1$.

Remark. Instead of in expectation, we can give similar bounds with high probability:

$$\mathbb{P}\left[\text{MSE}(\hat{f}) \leq C \cdot \text{MSE}(f_{\text{ORACLE}}) + T(n, D, f^*, K, d, \delta)\right] \geq 1 - \delta.$$

18.2.2 Oracle inequality for Ordinary Least Squares (OLS)

For OLS, we are given a matrix $\mathbf{X}_{n\times m}$ and we think of functions as $f_j(\mathbf{x}_i) = \mathbf{X}_{ij}$, where $\mathbf{x}_i$ is the i'th column of $\mathbf{X}$. Consider $\hat{f}_{\text{OLS}} = f_{\hat{\theta}^{\text{OLS}}}$.

Theorem 18.2 *Assume for each i, $\epsilon_i \in SG(\sigma^2)$. Then for $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$MSE(\hat{f}_{OLS}) \leq \inf_{\theta \in \mathbb{R}^m} MSE(f_\theta) + C\sigma^2 \left(\frac{m + \log(1/\delta)}{n}\right).$$

In the next lecture we prove this bound, and we consider such bounds for the Lasso.

References

[MW06] M. Wainwright, "Sharp thresholds for High-Dimensional and noisy recovery of sparsity using ℓ_1-Constrained Quadratic Programming (Lasso)," *IEEE transactions on information theory 55.5*, 2009, pp. 2183–2202.